DOCUMENT RESUME

ED 270 477 TM 860 358

AUTHOR Braungart-Bloom, Diane S.
TITLE Assessing Holistic Raters' Perceptions of Writing Qualities: An Examination of a Hierarchical Framework Following Pre-Post Training and Live Readings.
PUB DATE Apr 86
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986). For a related document, see TM 860 357.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS High Schools; *Holistic Evaluation; *Interrater Reliability; Judges; Measurement Techniques; *Pretests Posttests; Rating Scales; Scoring; Secondary School Teachers; State Departments of Education; State Programs; Student Evaluation; Testing Programs; *Training Methods; *Writing Evaluation; *Writing Skills
IDENTIFIERS *New Jersey Department of Education

ABSTRACT

The research described in this paper examines the relationship between raters and the training raters receive to apply that rating. The primary concern is maintaining scoring consistency within the context of raters' perceptions. If the criteria for holistic and/or analytic scoring can be recognized as communication based, rather than topic based, then student essays could be analyzed as features or qualities registered by the raters who read those essays. This paper examines 300 raters' perceptions of quality writing before training, after training, and following the reading of 90,000 student essays. Analyses of the holistic raters' perceptions of writing quality on a 1 to 6 score scale indicate 20 characteristics, grouped into six general constructs: (1) content; (2) organization; (3) sentence construction; (4) usage; (5) mechanical characteristics; and (6) textual characteristics. For pre-training scores the relationships among constructs remains consistent across ability levels. However, following training differentiation exists across all six categories, particularly at the extremes of the scales. Raters still seem less able to develop patterns of relationships among the essay characteristics at the middle points of the ability scales, even following training and reading of essays. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Assessing Holistic Raters' Perceptions of Writing Qualities: An Examination of a Hierarchical Framework Following Pre-Post Training and Live Readings

Diane S. Braungart-Bloom
Stanley Rabinowitz
New Jersey State Department of Education
Division of General Academic Education
Bureau of Academic and Cognitive Skills
225 West State Street
Trenton, New Jersey 08625

ED270477

ABSTRACT

Teachers participating in the holistic process of rating students' writing ability on a statewide examination have gained valuable insight into the skill of writing for specific students at a specific time. The research described in this paper examines the relationship between raters and the training raters receive to apply that rating. It examines raters' perceptions of writing before training, after training and following the actual reading of 90,000 students' essays. Furthermore, it identifies the hierarchical framework of writing qualities readers use of their knowledge base to guide them when applying a rating scale that assesses students' writing ability.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. S. Bloom

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Paper Prescribed at the 1986 Annual Meeting of the American Educational Research Association in San Francisco, California.

April 18, 1986

TM 860 358

Assessing Holistic Raters' Perceptions of Writing Qualities: An Examination of a Hierarchical Framework Following Pre-Post Training and Live Readings

Diane S. Braungart-Bloom
N.J. State Department of Education
225 West State Street
CN 500
Trenton, New Jersey 08625

DIANE S. BRAUNGART-BLOOM
STANLEY RABINOWITZ
New Jersey State Department of Education

Objective

Administrators of large-scale assessment programs are currently facing a variety of difficult problems in the area of writing which need to be resolved. Several states currently stipulate that in order for a student to graduate, that student must pass a statewide examination in writing to receive a state endorsed diploma. Because new forms of the tests are developed annually, scores on each examination must be equated or anchored from year to year to make certain that the passing score remains identical regardless of the test form administered. Equating the writing test's 30 minute multiple choice section is a relatively straightforward procedure (not unlike equating the multiple choice sections of reading and mathematics); however, the single generic prompt poses a more difficult resolution.

The research described in this paper examines the relationship between raters and the training raters receive to apply that rating. It examines raters' perceptions of writing before training, after training and following the actual reading of 90,000 students' essays. Furthermore, it identifies the hierarchical framework of writing qualities readers use as their knowledge base to guide them when applying a rating scale that assesses students' writing ability.

Perspective and Theoretical Framework

The rationale underlying holistic scoring is that it is "impressionistic" and that no aspect of writing skill can really be judged independently; however, if one can agree that raters can recognize and agree upon good writing when they see it, then it also stands to reason that raters can recognize those qualities that contribute to the strength and/or weakness of that piece of writing when it is being judged.

The primary concern of this study is maintaining scoring consistency within the context of raters' perceptions. If the criteria for holistic and/or analytic scoring can be recognized as communication based, rather than topic based, then student essays could be analyzed as features or qualities "registered" by the raters who read those essays. If such qualities can be "registered" and maintained year to year, equating can then be exercised because such qualities can be anchored by the very act of enforcing consistency on the salient qualities of a composition (organization, ideas, sentence structure, and mechanics).

Thus, recognizable and verifiable qualities of communication independent of topic and mode must be established to provide equivalent scoring from year to year. The concern lies in the fact that these "unknown" qualities may change because educator's instructional goals may have shifted from one year to the next (Eldridge, 1981). In an attempt to resolve this issue, a research study was developed based on examining raters' perceptions of "quality" writing before training, after training, and following the reading of 90,000 student essays.

Data Source/Techniques

A writing test was conducted for the second time in spring 1984 to 90,000 ninth grade students in New Jersey's public schools. The one-hour test measured students' ability to communicate effectively in writing (direct assessment) and apply certain mechanics to written text (indirect assessment). The 54-item multiple choice test measured sentence structure, organization of ideas, and editing. The 30-minute essay included a single generic prompt with pre-writing and editing. Essays, scored by New Jersey educators, used holistic analysis and a 6 to 1 scale (most to least competent).

Prior to the actual holistic training, prerequisite to rating the 90,000 essays, 300 educators were asked to complete a survey that asked them to rate the likelihood that various essay characteristics would appear in essays of varying quality. They were requested to use their own experience and expertise to consider the quality of typical essays that they would place at each point of a 6-point holistic scale (1=poor, 6=good). For each scale point educators rated how likely they thought essays assigned to that point would contain each of 20 characteristics listed. This survey was conducted prior to training, as well as following the training and reading of 90,000 essays rated over a 4-day time period.

For the purpose of this research, the pre and post results of 300 educators were obtained. The population was stratified to assure representation of knowledge and experience across the major types of geographic regions in New Jersey. For each educator, the following data were available--pre/post ratings; number of essays read; reliability; and percent essays read as discrepant.

Method and Results

Analyses of the Holistic raters' perceptions of writing qualities on a 1 to 6 score scale centered around readers' perceptions at the time of pre-treatment (before training) and at the time of post-treatment (following training and reading of 95,000 student essays.) This section discusses the results primarily in terms of the 20 characteristics as they were perceived by the readers at time 1 (prior to the training and rating sessions) and at time 2 (post condition ratings).

Eldridge, Richard "Grading in the 70s: How We Changed" College English 43 (1981), 64-68.

For each rater, both a pre- and post-training total score was computed for each of the six points of the score scale. The total score is an unweighted sum of the 20 essay characteristics appearing on the survey instrument. Separate analyses were performed on the pre and post scores. Each set of scores was factor analyzed by a form of Common Factor Analysis (Whittle, 1952) followed by an orthogonal rotation. Examination of the eigenvalues and residuals resulted in a two factor solution for the pre-training scores and a three factor solution for the post-training scores. Tables 1a and 1b should the resultant factor matrices.

TABLE 1

Factor Structure Matrices of Scale Points

a. Pre-Training

	Factor 1	Factor 2
Scale 1	-.120	.946
Scale 2	-.114	.957
Scale 3	.282	.900
Scale 4	.923	.214
Scale 5	.970	-.073
Scale 6	.944	-.123

b. Post-Training

	Factor 1	Factor 2	Factor 3
Scale 1	-.302	.912	.151
Scale 2	-.063	.775	.584
Scale 3	.239	.376	.872
Scale 4	.727	-.012	.652
Scale 5	.944	-.141	.206
Scale 6	.881	-.378	.136

As Table 1a indicates, raters prior to training and reading group the essay characteristics into two categories — low and high. Essays assigned a score of one to three are seen to possess similar attributes which differ from those rated four through six. Virtually no overlap appears between the two factors. Following training, however, greater differentiation appears. A third factor, representing a middle ability level is needed to account for that change. As Table 1b shows, the factor structure is no longer as "clear" as in the pre-training analysis. The raters appear to be using a more sophisticated classification scheme following training.

Factor scores were then computed to determine which of the 20 essay characteristics were responsible for the variation fround in the pre and post factor solutions.

As displayed in Table 2, raters prior to treatment (before training and reading of essays) have specific characteristics in mind that distinguish the six score points into two factor groupings — score points 1-2-3 (factor 1) and score points 4-5-6 (factor 2). The one characteristic constant across the two factor groups was the liklihood of an essay occuring in a narrative form. At score points 1-2-3 the narrative essay was rated as "very likely" to occur and at score points 4-5-6, a narrative essay was considered as "likely not" to occur. Of course, in terms of rhetorical structure, an expository essay was also considered "not likely" to occur.

TABLE 2

TIME 1 - PRE-TREATMENT

	SCORE POINTS	
RATINGS	1 - 3	4 - 6
Characteristics Rated Consistently Low (Likely to... not have (-) have (+)	(-) . figurative language (-) . paragraph structure/ transitions (-) . punctuation (-) . mechanics	(-) . narrative essay (+) . figurative language (-) . expository essay (-) . simple sentence structure
Characteristics Rated Consistently High (Likely to... not have (-) have (+)	(+) . narrative essay	(+) . focus (+) . a developed conclusion (+) . organized (-) . fragments/run-on sentences

In contrast, Table 3, exhibits significant changes in readers' judgements regarding the characteristics that distinguish the six score points. In Time 2 - Post-treatment (following training and reading of essays) raters were able to distinguish not just factor groupings across the 20 characteristics, but rather a 3 factor grouping of score points: 1-2; 3-4; and 5-6. Important to this study is that raters are aware of characteristics that affect a middle range of scores. However, what those characteristics are, is subject to confusion or disagreement by the raters. Some raters thought the characteristics were "likely" to be present at essays scored a 3 or 4, while other raters considered the characteristics as "not likely" to occur at those score points.

TABLE 3

TIME 2 - POST-TREATMENT

	SCORE POINTS		
RATINGS	1 - 2	3 - 4	5 - 6
Characteristics Rated Consistently Low (Likely to... not have (-) have (+)	(+) . narrative essay	(+) . narrative essay (+) . length	(+) . narrative essay (+) . expository essay
Characteristics Rated Consistently High (Likely to... not have (-) have (+)	(-) . organization (-) . spelling (-) . developed conclusion	(+)(-) . figurative language (+)(-) . parallel sentence structure (+)(-) . paragraph structure/ transitions (+)(-) . punctuation (+)(-) . mechanics (+)(-) . simple sentence structure	(+) . focus (+) . organized (+) . developed conclusion (-) . fragments/ run on sentences

Consistent throughout the pre and post-treatment was the low rating of rhetorical forms, i.e. narrative essays across the groupings. In addition, score point characteristics (narrative, expository) were again rated low following post-treatment, while figurative language and simple sentence structure were considered less important at post-treatment. Also evident at score points, 5-6 was the confirmation at post-treatment that focus, organization, and developed conclusion were "very likely" to be evident. Post-treatment did affect the ordering of these elements, however.

At the low score points, 1-2-3, pre-treatment rated the likelihood of narrative essays as an existing characteristic. Upon post-treatment, however, raters were more specific in what characteristics were "not likely" to be present - organization, spelling and developed conclusion were characteristics consistently rated as not likely to be present in essays rated as a 1 or 2 score point using the Holistic scoring method.

The 20 essay characteristics were grouped, based on expert ratings, into six general constructs -- Content, Organization, Sentence Construction, Usage, Mechanical Characteristics, and Textual Characteristics. This was done to increase the reliability of interpretations based on the ratings. At each of the six scale points, the general constructs were correlated among themselves,

for both pre and post ratings. This enabled us to examine the stability of the relationships among the six characteristics of essays, represented by the general constructs, along the overall writing ability continuum.

For the pre-training scores, the relationships among the constructs remained quite consistent across all the ability levels. The correlations were positive in all cases with the magnitude changing as a function of the number of individual essay characteristics making up the general construct. This indicates the differences are caused mostly by the stability of the composite each general construct represents.

However, this pattern changes following training. Much grater differentiation exists across the six categories, particularly at the extremes of the scales. Raters still seem less able to develop patterns of relationships among the essay characteristics at the middle points of the ability scales, even following training and reading of essays.

Educational Significance/Importance of the Study

The procedures for assessing students' writing ability have proven to be difficult and costly due in part to the paucity of research in this area and the ability to maintain scoring consistency for the purpose of eqating. The research described in this paper, then, examines the issues of scoring and the rating of essays that determine those qualities raters consider as important and necessary in order to provide large-scale assessment programs with the most valid and reliable measure of students' writing skills and ability. The New Jersey State Department of Education has responded to these issues with the development of a new scoring process intended for large scale assessment called The Registered Holistic Scoring Method (Bloom, D.S. 1986. NJSDE). The scoring method clarifies the difficult to distinguish points of the 1 to 6 score scale as it stabilizes and defines the composities of characteristics each construct represents at every score point.

DSB/kv:1/2606j